

Language Program Evaluation

Theory and practice

Brian K. Lynch

The University of Melbourne



CAMBRIDGE
UNIVERSITY PRESS

Published by the Press Syndicate of the University of Cambridge
The Pitt Building, Trumpington Street, Cambridge CB2 1RP
40 West 20th Street, New York, NY 10011-4211, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

© Cambridge University Press 1996

First published 1996

Library of Congress Cataloging-in-Publication Data

Lynch, Brian K.

Language program evaluation : theory and practice / Brian K.
Lynch.

p. cm. – (Cambridge applied linguistics series)

Includes bibliographical references (p.) and index.

ISBN 0-521-48191-0. – ISBN 0-521-48438-3 (pbk.)

1. Language and languages – Study and teaching – Evaluation.

I. Title. II. Series.

P53.63.L96 1996

418'.007 – dc20

95-1719

CIP

A catalog record for this book is available from the British Library

ISBN 0-521-48191-0 Hardback

ISBN 0-521-48438-3 Paperback

Transferred to digital printing 2003

Contents

Series editors' preface	ix
Preface	xi
1 Introduction	1
Definitions: Applied linguistics, evaluation, program	1
Critical issues	2
Program evaluation and applied linguistics research	9
2 Historical background	12
The paradigm dialog	12
Evaluation of language education programs in the 1960s and 1970s	22
More recent developments in the evaluation of language education programs	32
Summary	39
3 Validity	41
Validity from the positivistic perspective	41
Validity from the naturalistic perspective	53
Conclusions: Validity from the two perspectives	66
4 Positivistic designs	70
True experimental designs	71
Quasi-experimental designs	73
5 Naturalistic designs	80
The responsive model	80
The illumination model	82
Goal-free evaluation	84
The judicial model	85
The connoisseurship model	86
Other metaphors for naturalistic evaluation	88

viii *Contents*

6	Quantitative data gathering and analysis	92
	Data gathering	92
	Data analysis	94
	Conclusion	105
7	Qualitative data gathering and analysis	107
	Overview	107
	Data gathering	108
	Data analysis	139
8	Combining positivistic and naturalistic program evaluation	155
	Compatibilist versus incompatibilist perspectives	155
	Mixed strategies	156
	Multiple strategies	158
	Mixed designs	159
	Mixed designs and strategies over time	165
9	Conclusions	167
	CAM step 1 (audience and goals): Determine the purpose of the evaluation	167
	CAM steps 2 and 3 (context inventory and preliminary thematic framework): Determine what is being evaluated	170
	CAM steps 4 and 5 (evaluation design and data collection): Select a design and collect the data	171
	CAM step 6 (data analysis): Analyze and interpret your findings	173
	CAM step 7 (evaluation report): Communicating the evaluation findings	174
	The role of program evaluation in applied linguistics research	175
	References	178
	Author index	188
	Subject index	191

1 Introduction

It is probably safe to assume that the concept of program evaluation is not completely foreign to most applied linguists, even to those working outside the language education domain. Certainly the words *program* and *evaluation* conjure up reasonably clear mental images, and the notion that a program might need to be evaluated does not seem illogical to most. Language education programs abound internationally, and the majority of applied linguists have most likely, at some stage in their career, been involved in these programs as teachers, administrators, students, researchers, or some combination of these roles. Many, if not most, have been involved in some sort of effort to evaluate a language program. This evaluation may have taken the form of asking students to rate their language course and teacher using a questionnaire, giving achievement tests at the beginning and end of a period of instruction, or having a language teaching expert from another institution visit the program and prepare a report on its strengths and weaknesses. Program evaluation, then, can be seen as relevant to the experience of a wide range of applied linguists, and will be of particular interest to language educators.

Definitions: Applied linguistics, evaluation, program

In order to proceed with a detailed examination of the theory and practice of program evaluation within the broad context of applied linguistics, however, more precise definitions of certain terms are in order. I will focus on three key terms here; others will be presented in Chapter 2. To begin with, *applied linguistics* (AL), as a term defining an emerging academic discipline, has been the subject of recent discussions (Andersen et al. 1990; Pennycook 1990; van Lier et al. 1991). For the purposes of this book, AL will refer to research and practice concerned with the application of knowledge and methods from a variety of disciplines (e.g., anthropology, sociology, linguistics, psychology, and education) to the range of issues concerning the development and use of language.

2 *Language program evaluation*

The term *evaluation* tends to be used somewhat ambiguously in relation to other terms such as *assessment* and *testing*. Drawing upon the work of Bachman (1990) and Turner (1991), I will differentiate *evaluation* from these other terms primarily on the basis of its scope and purpose. That is, evaluation can make use of assessment instruments (including tests), but it is not limited to such forms of information gathering. It may include, for example, the use of unstructured interviews. Likewise, assessment instruments (including tests) can be used for purposes other than evaluation, such as to measure individual language ability in order to test a research hypothesis concerning language acquisition. Evaluation is defined here as the systematic attempt to gather information in order to make judgments or decisions. As such, evaluative information can be both qualitative and quantitative in form, and can be gathered through different methods such as observation or the administration of pencil-and-paper tests.

Program is a term that has perhaps been used with less ambiguity than evaluation. In general, it tends to evoke the image of a series of courses linked with some common goal or end product. A language education program generally consists of a slate of courses designed to prepare students for some language-related endeavor. This might mean preparing them to pass a language proficiency exam that, in turn, would allow them to gain entrance to some other program of study. It might also mean preparing them to function, in a general sense, in the context of a second language culture. These types of preparation can, of course, involve a single course (e.g., a Test of English as a Foreign Language [TOEFL] preparation course). In an effort to provide the broadest definition possible, I will use program to refer to any instructional sequence, such as a multilevel English as a second language (ESL) curriculum, a foreign language teacher-training workshop, a teaching unit being tried for the first time in a Japanese-for-business-purposes classroom, or computer-assisted instructional software that is self-accessed by students in a language lab.

Critical issues

The question that arises next, perhaps, is why applied linguists should concern themselves with program evaluation. In part, the answer to this question lies in the perennial need for language education programs to be evaluated, be it motivated by an internal quest for program improvement or by an externally imposed requirement in order to justify program funding. Accepting this, is program evaluation a generalized activity that has no need for a specific articulation within the context of applied linguistics? I believe that evaluation efforts do need to be

tailored to the specific concerns of language education programs (Lynch 1990b). Toward that end, I formulated the *context-adaptive model* (CAM) for language program evaluation (Lynch 1990a), drawing upon the historical development of program evaluation in applied linguistics that is discussed in the next chapter. Rather than a rigid model to be tested for validity using experimental research design and appropriate statistical techniques, it is meant to be a flexible, adaptable heuristic – a starting point for inquiry into language education programs that will constantly reshape and redefine itself, depending on the context of the program and the evaluation. I see the adaptable nature of the CAM as a partial antidote to many of the problems that have plagued previous attempts to evaluate language education programs. In the remainder of this introduction, I use this model as a framework for presenting the critical issues for language program evaluation. In addition to elucidating these issues, the CAM provides the basis for arguing for the important contribution that program evaluation can make to the development of applied linguistics as a field of research (see Figure 1.1).

Audience and goals

The first step of the CAM is concerned with identifying the audience and goals for the evaluation. Who is requesting the evaluation? Who will be affected by the evaluation? The answers to these questions help to determine the *stakeholders*, or *clients*, who have an immediate and central interest in the ultimate findings of the evaluation. Common examples of stakeholders are the program staff and the agencies that fund the program. The students of the program, although not clients per se, can also be thought of as an important audience for the evaluation. The concept of an evaluation audience can be broadened still further to include all those potentially interested in the conduct and results of the evaluation. Examples of this peripheral audience are program administrators, curriculum developers, teachers, and researchers from other program settings.

Identification of the evaluation audience leads to determining the evaluation goals, or purpose. Why is the evaluation being conducted? What information is being requested and why? Depending on the evaluation audience, the answers to these questions can be quite varied. Different subsets of the audience may also have different, and even conflicting, goals. For example, a funding agency may want statistical evidence that the program is producing higher test scores than some rival program, in order to justify continued financial support. The program staff, on the other hand, may want more descriptive information about how the instructional objectives are actually being realized in the classroom in order to improve the curriculum. These evaluation goals may

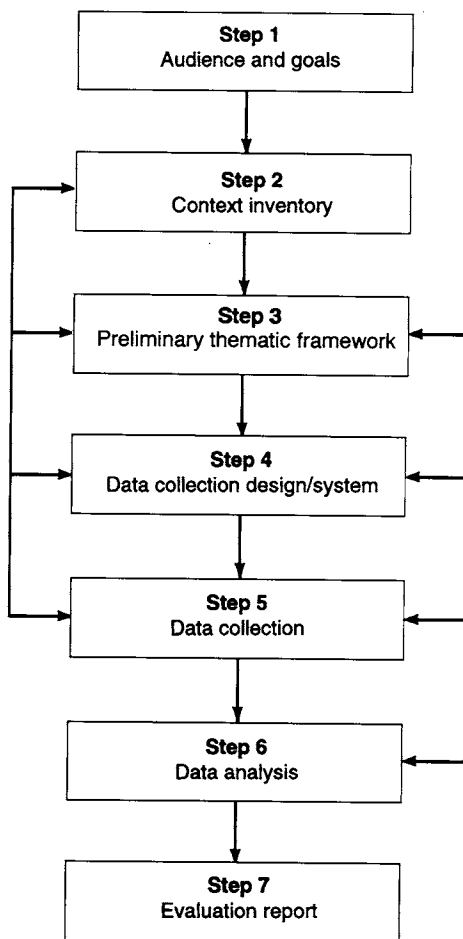


Figure 1.1 The context-adaptive model (CAM). (From B. K. Lynch 1990a:25. Copyright 1990 by TESOL. Reprinted by permission.)

have been stated clearly from the start by the evaluation audience, or they may need to be clarified by the evaluator in preevaluation interviews.

The particular evaluation audience and their goals for the evaluation will to a large extent determine the role of the evaluator. The funding agency, interested primarily in determining whether the program is “successful,” may require that the evaluation be carried out by persons external to the program, for greater objectivity. The program staff, interested in improving the curriculum, may elect an internal evaluation,

carried out by the program's own teachers and administrators, in order to take advantage of their close understanding of the program context. Of course, the decision to carry out an external versus internal evaluation will often be made based on the ability to find and pay for external evaluators. After being established as either external or internal, the role of the evaluator will be further defined, depending on the audience and goals (and, in certain cases, by the evaluator's preferred style), as someone providing consultation, an expert standing in judgment, a collaborator in program development, or a decision-making facilitator.

Context inventory

Another critical issue for program evaluation is the essential phenomena or features that characterize the program and its setting. The CAM addresses this issue with a checklist, or inventory, of potentially relevant dimensions of language education programs:

1. Availability of a *comparison group* (such as a "traditional" language program in a similar setting)
2. Availability of *reliable and valid measures* of language skills (criterion-referenced and/or norm-referenced tests, with program-specific and/or program-neutral content)
3. Availability of various types of *evaluation expertise* (such as statistical analysis, naturalistic research)
4. *Timing* of the evaluation (when the program begins, ends, and has breaks; how much time is available to conduct the evaluation)
5. The *selection process* for admitting students into the program (random selection, self-selection, selection according to preestablished criteria)
6. Characteristics of the program *students* (native language and culture, age, sex, socioeconomic status, previous education, previous academic achievement, previous experience with the language and culture being taught in the program)
7. Characteristics of the program *staff* (similar to characteristics of students; also, job descriptions, experience, availability, competence, and attitude toward the evaluation)
8. *Size and intensity* of the program (number of students, classrooms, proficiency/course levels, and number of hours per week/term)
9. *Instructional materials and resources* available to the program (textbooks, other instructional media and materials, human resources, basic office supplies)
10. *Perspective and purpose* of the program (notions, beliefs, and assumptions concerning the nature of language and the process of language learning; explicitly stated and informally articulated curricular goals)
11. *Social and political climate* surrounding the program (perception of the program by the surrounding academic and social community, student and community attitudes toward the language and culture being taught in the program, the relationship of the program's purpose to the larger social and political context)

Some of these dimensions will be more relevant in certain contexts than in others. Part of the adaptive nature of the CAM is the recognition that such an inventory will need to be tailored to the particular program setting. This tailoring may reflect practical constraints on the amount of detailed information capable of being gathered in the context inventory as well as the nature of certain dimensions (such as the unavailability of evaluation expertise or instructional materials and resources resulting from budgetary limitations). Along with the information on audience and goals, the context inventory acts as a guide for subsequent steps in the evaluation. It can act as an early indicator of the limits of a particular evaluation, and will inform decisions during subsequent steps in the evaluation process.

Preliminary thematic framework

The amount of information resulting from the first two steps of the CAM can be potentially overwhelming. A critical issue that arises at this early stage is how to focus the evaluation. Where should the evaluator begin? What aspects of the program should the evaluator investigate in detail? A preliminary thematic framework provides a conceptualization of the program in terms of the salient issues and themes that have emerged from the determination of audience and goals and the elaboration of the context inventory. Articulating this framework provides the evaluator with a focus that will guide the collection and analysis of evaluation data. The following is an example of a preliminary thematic framework developed for an English for science and technology (EST) reading program:

1. Effects of focusing instruction on reading only
 2. Effects of focusing instruction on reading skills and strategies
 3. Effects of using authentic reading texts
 4. Feasibility of using Spanish versus English for instruction
 5. Availability of classrooms
 6. Feasibility of using a "modified adjunct model" approach
 7. Feasibility and effects of conducting classroom-centered research
 8. Level of student proficiency in English upon entering the program
- (adapted from Lynch 1990a)

Data collection design/system

The evaluation audience and goals, context inventory, and preliminary thematic framework combine to suggest questions that the evaluator needs to answer. Another critical issue to be addressed is how best to obtain the information necessary to answer these questions. What type of data will need to be gathered – quantitative, qualitative,

or both? What will be the best methods for gathering the data? If the primary evaluation question is “Are the students of this program making significant gains in their language abilities?” then a quantitative design may be most appropriate. Language ability test scores would be gathered before and after participation in the program and analyzed for statistical significance. On the other hand, if the primary question to be answered is “How can we improve this program?” then a qualitative design may be called for. The evaluator(s) would observe program classes, interview students and staff, and try to describe how the program is functioning in order to make recommendations for change. In other evaluation contexts, there may be a combination of questions to answer that require both quantitative and qualitative data.

The context inventory is extremely useful at this stage for determining the feasibility of certain types of data collection design. In particular, the lack of availability of a comparison group will severely constrain the range of quantitative designs that are possible. That is, without a comparison group, the evaluation will not be able to make use of most traditional experimental and quasi-experimental research designs. Dimensions such as the attitude of program staff toward the evaluation and the availability of evaluation expertise will also dictate limits for the data collection design. For example, an unwillingness on the part of the program staff to provide time for evaluation efforts or the lack of available expertise in qualitative data analysis (the evaluator may be untrained in this type of evaluation and may be unable to procure such expertise) may result in the evaluator abandoning plans to collect interview data from program participants. Two examples of the interaction of audience and goals, context inventory, and preliminary thematic framework in the selection of a data collection design are represented in Figure 1.2.

Data collection and analysis

Data collection and analysis follow logically from the type of design chosen for the evaluation. The critical issues that concern the evaluator here have to do with the appropriate conduct of the data-gathering procedures and the interpretation of the results. In the case of quantitative designs, have the assumptions of the design and statistical models been met? In the case of qualitative designs, have the procedures for data gathering been portrayed accurately, and have alternative interpretations of the data been pursued? Like the choice of evaluation design, these are obviously complex issues that are discussed in greater detail in subsequent chapters.

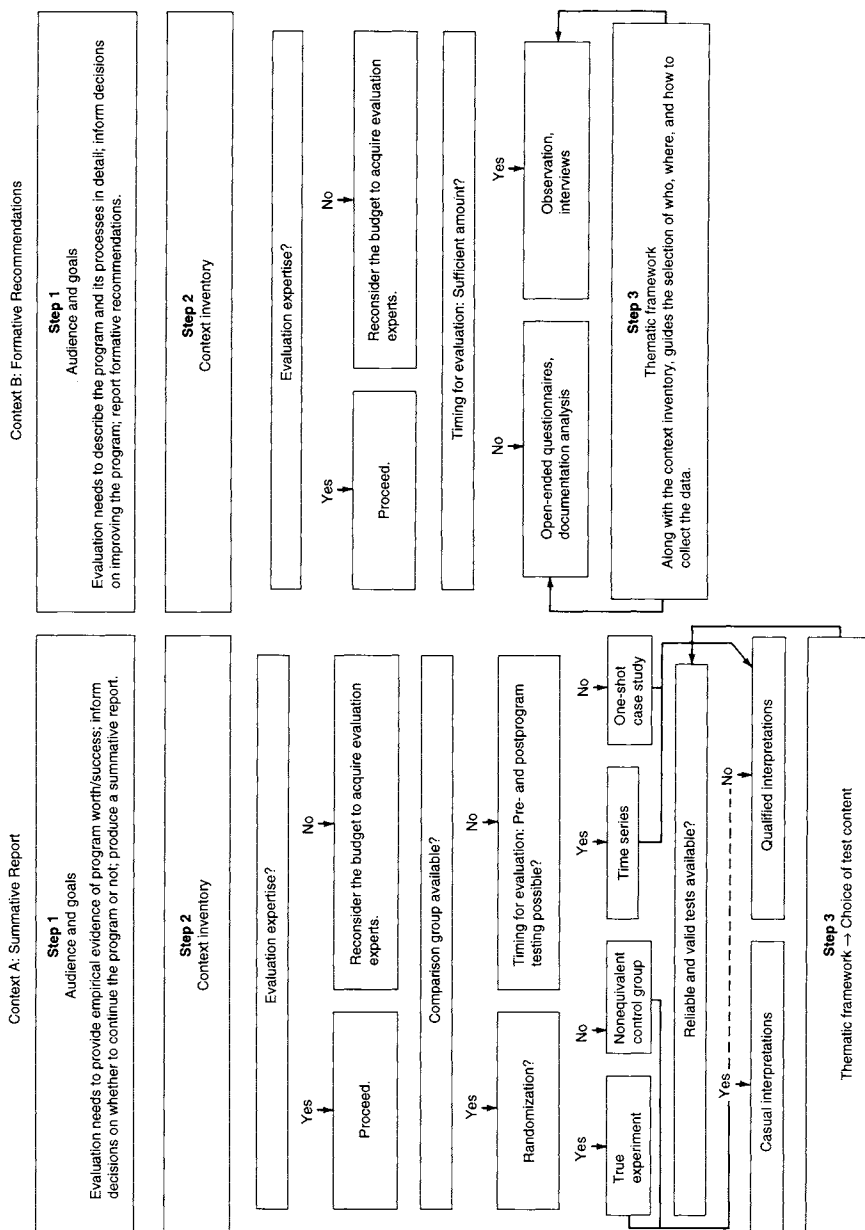


Figure 1.2 *Formulation of the data collection design/system. (From B. K. Lynch 1990a:34-35. Copyright 1990 by TESOL. Reprinted by permission.)*

Evaluation report

In order to produce a useful final report, the evaluator must be extremely sensitive to the audience and goals of the evaluation. The social and political climate dimension of the context inventory needs to be considered carefully at this stage as well. The critical issue is how to communicate the findings of the evaluation honestly and successfully. The evaluator may find that certain conclusions and the evidence on which they are based need to be omitted from this communication, be it a formal, written document or an informal, oral report. Rather than a covering up or withholding of the truth, this should be seen as a concern for communicating effectively with the audience. Topics that are extremely sensitive or issues that are tied to specific personalities may obscure the intended message and lead to a misunderstanding of the evaluation's conclusions and the evidence on which they are based. The evaluator may find it necessary to provide multiple reports that highlight different types of evaluative information or that express the evaluation findings in different ways, depending on the intended audiences.

Program evaluation and applied linguistics research

I have argued elsewhere that program evaluation can play an essential role in the development of applied linguistics as a field of research (Lynch 1991). Given the critical issue of identifying the audience and goals for an evaluation discussed in the previous section, the work of program evaluation leads to a careful consideration of what counts as evidence. In the literature on program evaluation this has been characterized as the *qualitative-quantitative debate* (Reichardt and Cook 1979; Smith and Heshusius 1986; Howe 1988; Smith 1988), *paradigm wars* (Gage 1989), or *paradigm dialog* (Guba 1990). This debate is summarized in Chapter 2 as part of the historical background for program evaluation in applied linguistics. The point to be made here, however, is that there exists work that has brought knowledge concerning research paradigms from other disciplines to bear on the evaluation of language education programs (Long 1983; Beretta 1986a; Lynch 1988, 1992) and thus adds to the definition of applied linguistics as a field of research. This work has raised the important issue of what we accept as evidence for answering our evaluation questions. Different audiences for a program evaluation force the evaluator to consider the issue of what counts as evidence from different perspectives. As mentioned previously, a funding agency may expect statistics as proof that a program deserves continued support. The program staff as audience for an evaluation may expect a clear description of how the program is actually

functioning. Evaluators need to consider these perspectives along with their own requirements for what counts as evidence. I believe that this process strengthens research in applied linguistics by opening the field to different types of knowledge and knowledge validation and by clarifying the bases on which it is built.

In addition to keeping us aware and honest about what counts as evidence in our inquiry, program evaluation can spark investigation across a wide range of research areas that describe applied linguistics. The dimensions of the context inventory, discussed previously as a checklist for gathering the necessary information about a program, involve consideration of such issues as the social and political basis and motivation for language learning and teaching. The concern for reliable and valid measures leads program evaluators into the area of language testing and the application of knowledge and techniques from education and psychology for the improved measurement of language ability. When the evaluation questions to be answered involve a combination of "Has it succeeded?" and "How has it succeeded?" a multiple-research-methods strategy that leads program evaluation into complex qualitative-quantitative designs is called for. For example, in order to investigate the match between program objectives and classroom processes, an evaluator needs to consider a combination of precise quantitative measurement of student achievement and qualitative methods such as an ethnographic description of the classroom, in addition to introspective/retrospective investigation of individual learning processes and their interaction with instruction. If the instructional objectives of a program are based on second language acquisition (SLA) theory, program evaluation can provide a testing ground for SLA research.

What counts as evidence? What are the social and political factors that affect language learning and teaching? How can we best define and measure language abilities? What are the best research designs for our inquiry? How do learners acquire a second language? These questions, which cut a wide swath across the applied linguistics terrain, are all critically related to the enterprise of program evaluation. The chapters that follow attempt to lay the theoretical foundations of this enterprise, as well as provide the practical means for its conduct.

Chapter 2, as mentioned, outlines the differences between the competing research paradigms in program evaluation (the quantitative-qualitative debate). Ultimately, I argue for the use of both paradigms, which I refer to as *positivistic* and *naturalistic*, while maintaining an awareness of the epistemological differences that divide them. Following this discussion of research methodology, I present a history of language education program evaluation from the 1960s to the present. The central focus of this presentation is the shift from essentially quantitative, positivistic studies that look only at end-of-program achievement gains

to ones that include an investigation of program process using qualitative, naturalistic methods.

Chapter 3 discusses the issue of validity from both the positivistic and naturalistic perspectives. First I present the classical notions of internal and external validity, as well as some more recent approaches within the positivistic paradigm. Validity within the naturalistic paradigm is then discussed, highlighting the fundamental similarities to and differences from the positivistic paradigm.

Chapter 4 explains the various models or research designs for evaluation within the positivistic paradigm. I contrast true experiments with quasi-experimental design, and discuss the issues of control (or comparison) groups, selection, and measurement.

Chapter 5 presents a variety of models for carrying out program evaluation within the naturalistic paradigm. I present various examples of how certain “metaphors” for evaluation (Smith 1981) might be defined in the language teaching context.

Chapter 6 presents various techniques for collecting and analyzing quantitative data. First, I discuss the issue of the appropriate quantitative instruments, including consideration of norm-referenced versus criterion-referenced measurement and the selection of test content. Using example data from the Reading English for Science and Technology (REST) project evaluation (Lynch 1988, 1992), I present a variety of statistical procedures for analyzing quantitative data. This presentation is accompanied by a discussion of the requirements for using and interpreting the various statistical models in the context of program evaluation.

Chapter 7 begins with a discussion of qualitative data-gathering techniques. It focuses on observation and interviewing, but also considers such techniques as document analysis and journal keeping. I then present techniques for reducing the qualitative data, analyzing it, and forming interpretations. These techniques are illustrated with examples from the REST project evaluation (Lynch 1988, 1992).

Chapter 8 gives examples of program evaluation models that combine features of the positivistic and naturalistic approaches. I discuss various ways of mixing qualitative and quantitative data, analytic techniques, and designs.

In the final chapter, I summarize the theoretical and practical issues presented in the preceding chapters, focusing on the various purposes, contexts, designs, analyses, and modes of reporting results. Finally, I review the potential role for program evaluation in applied linguistics research.